A physiologically inspired model of auditory stream segregation based on a temporal coherence analysis

Christiansen, Simon Krogholt; Jepsen, Morten Løve; Dau, Torsten

Published in:
Proceedings of Meetings on Acoustics

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Proceedings of Meetings on Acoustics

Volume 15, 2012 http://acousticalsociety.org/

163rd Meeting Acoustical Society of America/ACOUSTCS 2012 HONG KONG

Hong Kong
13 - 18 May 2012

Session 1aPP: Psychological and Physiological Acoustics

1aPP8. A physiologically inspired model of auditory stream segregation based on a temporal coherence analysis

Simon Krogholt Christiansen*, Morten L. Jepsen and Torsten Dau

***Corresponding author's address: Centre for Applied Hearing Research, Technical University of Denmark, Kgs. Lyngby, 2800, -, Denmark, skch@elektro.dtu.dk**

The ability to perceptually separate acoustic sources and focus one's attention on a single source at a time is essential for our ability to use acoustic information. In this study, a physiologically inspired model of human auditory processing [M. L. Jepsen and T. Dau, J. Acoust. Soc. Am. 124, 422-438, (2008)] was used as a front end of a model for auditory stream segregation. A temporal coherence analysis [M. Elhilali, C. Ling, C. Micheyl, A. J. Oxenham and S. Shamma, Neuron. 61, 317-329, (2009)] was applied at the output of the preprocessing, using the coherence across tonotopic channels to group activity across frequency. Using this approach, the described model is able to quantitatively account for classical streaming phenomena relying on frequency separation and tone presentation rate, such as the temporal coherence boundary and the fission boundary [L. P. A. S. van Noorden, doctoral dissertation, Institute for Perception Research, Eindhoven, NL, (1975)]. The same model also accounts for the perceptual grouping of distant spectral components in the case of synchronous presentation. The most essential components of the front-end and back-end processing in the framework of the presented model are analysed and future perspectives discussed.

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.4729595]
Received 24 May 2012; published 13 Jun 2012

INTRODUCTION

In a natural acoustic environment the sound that reaches our ears is a complex mixture of sound sources. In order to parse this stimulus, our central auditory system segregates the signal into separate auditory objects or "streams" [1]. This allows us to selectively attend to a single auditory stream, and thus to focus on a conversation or piece of music while ignoring competing acoustic information. Models of auditory stream segregation that rely on frequency or tonotopic separation of the sound components for stream segregation have been proposed (e.g. [2]). Physiological studies in animals have also used tonotopic separation as a measure of stream segregation and have shown a correlation between tonotopic separation and psychophysical stream segregation (e.g., [3]). While tonotopic separation may be necessary for stream segregation, it cannot explain the grouping or fusion of distant spectral components due to e.g. common pitch or onset/offset synchrony. These cues allow fusion of sounds into the same perceptual stream despite tonotopic separation [4, 5].

In the present study, a physiologically inspired model of auditory stream segregation is presented. The model is based on the computational auditory signal-processing and perception (CASP) model [6] combined with the conceptual model of grouping suggested by Elhilali et al. [4]. The proposed model was applied to two different stream segregation experiments. The first experiment investigated the model's ability to account for grouping of distant spectral components due to synchrony; specifically, how varying the degree of asynchrony affects the grouping of otherwise isochronous tone sequences. The second experiment investigated the model's ability to account for the stream segregation observed by van Noorden [7], where a "galloping" tone pattern was presented producing different streaming percepts depending on the tone rate and the frequency separation of the tones. For this experiment, the model results are compared to the results from van Noorden [7].

MODEL FRAMEWORK

The model consists of two parts: A decomposition stage (peripheral processing and modulation filtering) based on CASP, and a grouping stage (temporal coherence analysis) based on the conceptual model by [4].

Peripheral Processing and Modulation Filtering

The peripheral processing stage consists of a basilar-membrane filterbank, a hair-cell transduction stage, and an adaptation stage. The basilar membrane filterbank is implemented as a 4th order gammatone filterbank [8] with one equivalent rectangular bandwidth (ERB) [9] spacing. The hair-cell transduction stage is realized by half-wave rectification followed by low-pass filtering at 1 kHz. Neural adaptation is modelled by five feedback loops connected in series with time-constants ranging from 5 to 500 ms [10].

The output of the peripheral stage is processed by a first-order low-pass filter with a cut-off frequency of 150 Hz, simulating the decreasing sensitivity to sinusoidal modulation as a function of modulation frequency. The low-pass filter is followed by a modulation filterbank. This is functionally similar to the temporal integration stage used in [4]. The modulation filterbank consists of band-pass filters with center frequencies ranging from 0 (low-pass filter) to 1000 Hz [6].

Coherence Analysis

Stream segregation is determined based on correlation between auditory channels. Channels with positively correlated activity over time are assigned to the same perceptual stream. As in Elhilali et al. [4], a windowed correlation between each pair of peripheral channels is computed by multiplying each pair of modulation filtered peripheral channels. The result is presented as a dynamic coherence matrix that shows the correlation between the peripheral channels over time.

To quantify the coherence matrix, an eigenvalue decomposition is performed. The decomposition shows channels that are positively correlated with each other (and form a stream). The eigenvalue decomposition determines the number of independent dimensions of the coherence matrix, and by analogy, the number of streams present in the stimulus [4].

In the current study, it is of interest whether a stimulus is perceived as one or two streams, and thus, whether there are one or two significant eigenvalues. The ratio of the second largest eigenvalue (λ_2) to the largest eigenvalue

(λ_1) is therefore used as a measure of the "strength" of the two-stream percept. If the coherence matrix can be decomposed into one main component, the ratio λ_2/λ_1 will be very low (close to zero), corresponding to a one-stream percept. If the ratio λ_2/λ_1 is high, this indicates that there are (at least) two significant dimensions, and thus, at least two streams.

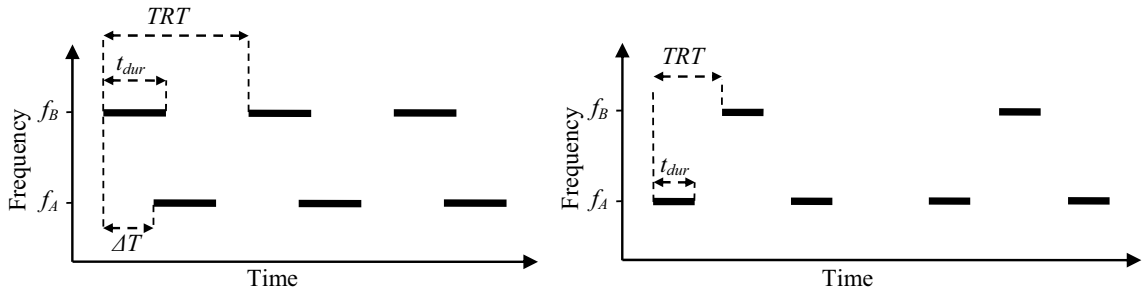
METHOD

Experiment I – Grouping of Distant Spectral Components due to Synchrony

A sketch of the stimuli is presented in Figure 1a. The stimuli consisted of two repeating pure tones, A (f_A = 300 Hz) and B (f_B = 952 Hz, 20 semitones higher) with duration t_{dur}. Each tone was gated on and off using 10 ms raised cosine ramps. Tone pairs were repeated at a fixed tone repetition time (TRT). For each tone pair, the onset of the B occurred ΔT after the onset of A. Over the course of a sequence ΔT was varied linearly, starting at -100 ms (B leading A) and ending at +100 ms (B lagging A). Six experimental sequences were constructed from combinations of two tone durations (t_{dur}; 30, 75 ms) and three tone repetition times (TRT; 150, 200, 250 ms). The gradual change of ΔT was implemented by using a longer TRT for the B-tones (150.4, 200.5, 250.6 ms). Each combination of TRT and t_{dur} was presented 5 times, resulting in 30 measurements per test subject. The entire experiment was repeated with ΔT changing in the opposite direction (from +100 to -100 ms), to avoid a response bias due to the direction of change of ΔT.

Experiment II – Stream Segregation of Alternating Tones due to Frequency Separation and Tone Repetition Time

A schematic representation of the stimuli is shown in Figure 1b. The stimuli consisted of two pure tones, A and B, presented in an ABA-ABA pattern. Each tone was 40 ms long and gated on and off using 5 ms raised cosine ramps. The onset-to-onset time between alternating tones was controlled by TRT. The frequencies of tones A and B were set to 1 kHz and *N* semitones higher. Combinations of ten TRT values (60 - 150 ms in steps of 10 ms) and 31 levels of *N* (0 – 15 semitones) were tested with the model (310 different conditions in total). The eigenvalue ratio λ_2/λ_1 was calculated for each condition.



(a) – Experiment I (b) – Experiment II

FIGURE 1. Schematic representation of the stimuli used in Experiment I (a) and Experiment II (b)

Apparatus and Test Subjects

All stimuli were generated in Matlab (44.1 kHz sampling rate, 24-bit precision; Mathworks) and presented over headphones (Sound card: RME DIGI 96/8 PAD; Headphones: Sennheiser HD580).

The participants were three normal-hearing listeners ranging between 23 and 26 years of age. All listeners were trained in each experimental condition until their results stabilized. The subjects were seated in front of a computer monitor in a double-walled, sound-insulated booth.

Two buttons were presented on the computer screen. At the start of the experiment, the "2 streams" button was activated. The subjects were instructed to push the appropriate button when the tones fused into a single stream or a single "sound event" or split. Stream fusion/segregation thresholds were estimated from the ΔT at the times the button presses were recorded.

RESULTS

Experiment I - Grouping of Distant Spectral Components due to Synchrony

The stream fusion/segregation thresholds from Experiment I are shown in Figure 2a. A 3-way ANOVA was conducted with three factors: TRT, t_{dur}, and direction of asynchrony (sign(ΔT)). While the main effect of direction of asynchrony was significant [$F(1,24)=4.45$, $p<0.05$], no other main effects or interactions were significant.

Results from the model using the same stimuli are plotted in Figure 2b. The lines indicate contours with fixed eigenvalue ratios. Like the behavioural results, the model shows only a minor effect of TRT and t_{dur}. The simulation results do, however show an asymmetry of the one-stream percept, favouring a lagging B-tone (a positive ΔT).

Experiment II - Stream Segregation of Alternating Tones due to Frequency Separation and Tone Repetition Time

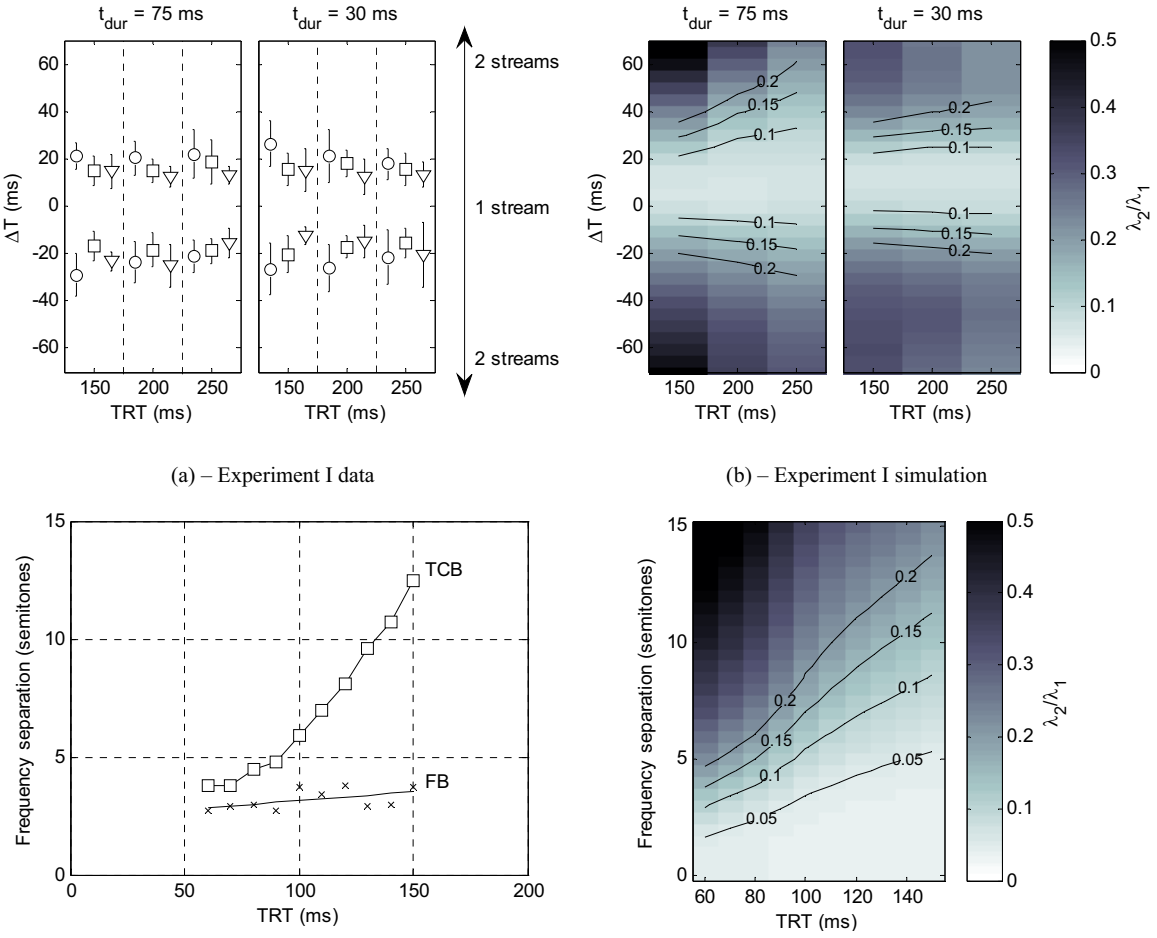
Data from a previous behavioural experiment by van Noorden [7] are shown in Figure 2c and results from our model are plotted in Figure 2d. In panel c, the curves indicate the temporal coherence boundary (TCB) and the fission boundary (FB). Above the TCB, the stimulus is always perceived as two streams, and below the FB the stimulus is always perceived as one stream. In the range between the curves, the percept can be controlled and either a fused or segregated percept can be achieved. In panel d, the lines indicate contours with fixed eigenvalue ratios.

DISCUSSION

The results of Experiment I demonstrates that temporal synchrony facilitates perceptual fusion of tones despite large frequency separations. These results were compared with our proposed model that segregates an auditory signal into streams if the activity across the output of neural channels is incoherent. In experiment I, the frequency separation between the two tones was fixed at 20 semitones and the asynchrony of the tones was varied over time. Thus, each tone produced activity in different neural channels and the coherence of the activity across channels varied. By choosing an appropriate threshold of coherence, the model is able to simulate the behavioural data quite well.

Bregman [1] investigated the ability of two pure tones to fuse into a single stream based on the tones' relative synchrony. His study used identical TRTs for the two tones, similarly to our study. His study showed that the tones would segregate into separate streams if the temporal overlap was less than 50 %, and fuse into a single stream if the overlap exceeded 88 %. This is in contrast with our study, where a ΔT of 10-30 ms was observed for the different subjects, with no significant effect of either TRT or t_{dur}. This means that the long duration tones (t_{dur} = 75 ms) fused into a single stream when the temporal overlap exceeded 60-87 %, and the short duration tones (t_{dur} = 30 ms) fused into a single stream even without any temporal overlap (ΔT = 30 ms). Reconsidering the study of [1], only a single tone duration was used (t_{dur} = 250 ms), and converting the 88 % overlap into an absolute asynchrony yields a ΔT = 30 ms, which corresponds fairly well with the results from our study.

In our model results, an asymmetry with respect to ΔT was observed. This asymmetry is caused by the bandwidths of the gammatone filterbank and their associated time-constants. The gammatone filters of the peripheral filterbank with center-frequencies closest to the A and B tones have time constants of 17.6 and 7.4 ms, respectively. Due to this frequency dependent delay, the maximum coherence at the output of the peripheral stage occurs when tone A leads tone B by approximately 10.2 ms, causing the asymmetry observed in Figure 2b. The data also show a small (but significant) effect of the direction of the delay ΔT, but in the other direction (mean ΔT = {-20.2, 16.9} ms).



(a) – Experiment I data

(b) – Experiment I simulation

(c) – Experiment II data from [7].

(d) – Experiment II simulation

FIGURE 2. Results from Experiment I (a) and (b) and Experiment II (c) and (d). The symbols in (a) show the mean asynchrony ΔT indicating the transition between 1 and 2 streams. The different symbols represent different test subjects. Part (c) show experimental results from [7]. The upper curve shows the temporal coherence boundary (TCB), and the bottom line shows the fission boundary (FB). The simulations of the two experiments are shown in (b) and (d) for Experiment I and II, respectively. The grey scale intensity indicates the eigenvalue ratio. A bright colour represents a low eigenvalue ratio, corresponding to a one-stream percept, and a dark colour represents a high ratio, corresponding to a two-stream percept. The curves indicate contours with fixed eigenvalue ratios.

The difference in asymmetry from model results to experimental data may be caused by higher-level processing that compensates for the frequency dependent delays of the basilar-membrane filtering, as indicated in a recent study by Wojtczak et al. [11]. Their study investigated synchrony detection of pure tones with varying frequency separations, and found no effect of basilar membrane latency. On the contrary, their experiments showed an asymmetry favouring low-frequency components lagging high-frequency components. These findings are consistent with our study, which indicates that the perceptual grouping of stimuli takes place at a later stage in the auditory system than the compensation for basilar membrane latency.

In the second experiment, the measured (and simulated) data show that fast repeating tone sequences are more likely to split into two separate streams, whereas slowly repeating tone sequences can be perceived as a single stream for much larger frequency separations. In the modeling framework, a two-stream percept only occurs in the situation where (at least) two channels contain incoherent activity. Thus, in order to produce a two-stream percept, the stimuli must at least activate two separate peripheral filters. For the lowest non-zero frequency separation used in the simulation, the A and B tones have frequencies of 1 kHz and 1.06 kHz. The bandwidth of the gammatone filter

centered at 1 kHz is 133 Hz and both tones will in this case be processed by the same peripheral filters. Therefore, the model does not predict a two-stream percept. At larger frequency separations, e.g. 7 semitones, a substantial difference in the model results is observed between low and high TRTs. Since the frequency separation is the same, this cannot be explained by effects of spectral separation. Instead, the different results are caused by the adaptation stage in the peripheral model which accounts for forward masking. The forward masking effect reduces the sensitivity of a peripheral channel after a tone has been presented, which effectively reduces the spread of excitation of the other tone. This, in turn, reduces the temporal coherence of the channels, causing the model to predict a two-stream percept.

Physiological studies on animals [3, 12] also suggest physiological forward masking as a possible cause of the stream segregation observed in the experimental paradigm utilized by van Noorden [7]. For tone sequences with small frequency separations, both tones in the stimuli were able to excite an auditory nerve tuned to one of the two frequencies. When the tone rate was increased (lower TRT), the excitation from the non characteristic-frequency tones was reduced, resulting in a reduced coherence of the auditory nerves tuned to the two frequencies. The observed behaviour corresponds to forward masking, with suppression of neural responses to a sound (the signal) following the presentation of a preceding sound (the masker). The results from our study support this hypothesis.

SUMMARY AND CONCLUSION

The present study presented a physiologically inspired model of auditory stream segregation. The premise of the model is that temporal coherence determines the perceptual organization of stimuli, by fusing coherent channels into a single stream. As a consequence, within the model, stream segregation can only occur if a stimulus is able to evoke non-coherent activity in different channels of the model. This requires tonotopic separation and temporal incoherence of spectral components. The tonotopic separation can either be achieved through a sufficiently large frequency separation, or, as in the case of Experiment II, as a consequence of forward-masking.

The presented model is based on a physiologically inspired model of the peripheral auditory system [6], which is able to account for a range of psychophysical phenomena. By the simple addition of the temporal coherence analysis [4], the model was shown to account for both classical stream segregation experiments [7], as well as spectral fusion of stimuli due to synchrony.

The functionality of the model is currently limited to providing an estimate of the “strength” of the two-stream percept. It cannot actually segregate sound sources. However, the suggested model may help to understand the underlying processes involved in primitive stream segregation, as was demonstrated in Experiment II with the influence of forward-masking on the perceptual organization of the tones.

Furthermore, the model is purely bottom-up based, and cannot utilize cues that require top-down processing, such as recognition of known sounds, etc. At the present stage, the model is also purely monaural, and can thus not utilize spatial cues, which are known to play a role in segregating sound sources [1].

REFERENCES

1. A. S. Bregman, *Auditory scene analysis*, Cambridge, MA-MIT Press, (1990).
2. S. McCabe, and M. J. Denham, “A model of auditory streaming,” *J. Acoust. Soc. Am.* **101**, 1611-1621 (1997).
3. M. A. Bee, and G. M. Klump, “Auditory stream segregation in the songbird forebrain: effects of time intervals on responses to interleaved tone sequences,” *Brain Behav. Evol.* **66**, 197-214 (2005).
4. M. Elhilali, C. Ling, C. Micheyl, A. J. Oxenham, and S. Shamma, ”Temporal coherence in the perceptual organization and cortical representation of auditory scenes,” *Neuron* **61**, 317-329 (2009).
5. C. Micheyl, C. Hunter, and A. J. Oxenham, “Auditory stream segregation and the perception of across-frequency synchrony,” *J. Exp. Psychol. Hum. Percept. Perform.* **36**, 1029-1039 (2010).
6. M. L. Jepsen, S. Ewert, and T. Dau, “A computational model of auditory signal processing and perception, “*J. Acoust. Soc. Am.* **124**, 422-438 (2008).
7. L. P. A. S. van Noorden, “Temporal coherence in the perception of tone sequences,” doctoral dissertation, Institute for Perception Research, Eindhoven, The Netherlands (1975).
8. R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, “An efficient auditory filterbank based on the gammatone function,” in paper presented at a meeting at the IOC Speech Group on Auditory Modelling at RSRE, December 14-15 (1997).
9. B. R. Glasberg, and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hear. Res.* **47**, 103-138 (1990).

10. T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the 'effective' signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Am.* **99**, 3615-3622 (1996).
11. M. Wojtczak, J. A. Beim, C. Micheyl, and A. J. Oxenham, "Perception of across-frequency asynchrony and the role of cochlear delays," *J. Acoust. Soc. Am.* **131**, 363-377 (2011).
12. Y. I. Fishman, J. G. Arezzo, and M. Steinschneider, "Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration," *J. Acoust. Soc. Am.* **116**, 1656-1670 (2004).